DOI: https://doi.org/10.70656/ijcse.v3i01.661

A Comparative Study of SMOTE Variants and Particle Swarm Optimization for Feature Selection and Hyperparameter Tuning on Decision Tree in Breast Cancer Classification

Himam Bashiran
Faculty of Informatics
Telkom University
Purwokerto, Indonesia
himamb@student.telkomuniversity.ac.id

Fito Satrio
Faculty of Informatics
Telkom University
Purwokerto, Indonesia
fitosatrio@student.telkomuniversity.ac.id

Agung Malik Ibrahim
Faculty of Informatics
Telkom University
Purwokerto, Indonesia
agungibr@student.telkomuniversity.ac.id

Abstract*—Breast cancer is one of the most prevalent types of cancer in Indonesia, making accurate early detection highly crucial to reduce mortality rates. However, classification effectiveness is often hindered by high-dimensional data and class imbalance issues, which can introduce bias into predictive models. This study proposes the integration of the Decision Tree algorithm with hybrid sampling techniques and Particle Swarm Optimization (PSO). PSO is employed to perform a dual role, namely simultaneous feature selection and hyperparameter tuning. Experimental results show that using 30 particles in PSO successfully reduced the data dimensionality significantly from 30 features to only 6 essential features: texture1, symmetry1, radius2, area3, smoothness3, and symmetry3. The combination of SMOTE-Tomek and PSO-based feature selection emerged as the best-performing scenario, achieving an accuracy of 99.68% while producing only one False Negative prediction. In addition to superior precision, the model demonstrated remarkable computational efficiency with an average latency of 0.0036 ms and a throughput of 274,897 samples per second. Explainable AI (XAI) analysis using SHAP confirmed that area3 was the most dominant feature, which is clinically consistent with indicators of cancer cell proliferation. This study proves that the synergy between data balancing techniques and metaheuristic optimization can produce an accurate, transparent, and efficient diagnostic model suitable for real-time medical implementation.

Keywords- Breast Cancer, Decision Tree, Particle Swarm Optimization, SMOTE-Tomek, Feature Selection, Hyperparameter Tuning

I. INTRODUCTION

Cancer is a group of diseases characterized by the uncontrolled growth of abnormal cells that have the potential to spread to other parts of the body [1]. Among its various types, breast cancer is one of the most dangerous non-skin cancers in women and ranks as the second leading cause of cancer-related deaths worldwide after lung cancer [2]. This disease generally develops in the ductal epithelium or lobules of the breast and is caused by various pathological factors, ranging from abnormalities in cells and ducts to the supporting tissues of the breast, excluding the breast skin [3], [4].

In Indonesia, data from the Global Burden of Cancer (IARC) indicate that breast cancer has the highest incidence of new cancer cases [5]. Despite its high level of severity, the progression of breast cancer is often asymptomatic or does not show symptoms in its early stages, so most new cases can only be detected through routine examinations [6]. Therefore, early detection and accurate classification are highly crucial to reduce patient mortality rates.

The utilization of Machine Learning techniques has made significant contributions in assisting medical professionals to diagnose breast cancer objectively. Several previous studies have examined the effectiveness of various algorithms in determining cancer types, whether malignant or benign tumors. For example, the use of the Support Vector Machine (SVM) algorithm demonstrated a high accuracy rate of 94% [7]. In addition, the K-Nearest Neighbor (KNN) algorithm has also been implemented for breast cancer classification with an accuracy of 93% [3]. Another study comparing the performance of Random Forest and KNN in diagnosing this disease achieved an accuracy of 91.18% [8].

Along with technological advancements, research has shifted toward the use of hybrid methods to improve classifica-tion efficiency through optimization techniques. One promis-ing approach is the integration of Particle Swarm Optimization (PSO) with Decision Tree. The use of PSO in this context has proven effective in handling high-dimensional medical data through a feature selection mechanism, achieving an accuracy of up to 92.26% [9]. This optimization process is crucial for eliminating irrelevant attributes so that the model becomes more efficient and accurate.

However, the effectiveness of optimized classification mod-els is often hindered by the characteristics of imbalanced medi-cal datasets. This imbalance becomes a fundamental challenge because machine learning algorithms tend to overclassify the majority class, resulting in high accuracy for the majority class but very low performance for the minority class [10]. This issue is highly critical in medical diagnosis. For example, in lung cancer classification, the use of a model without handling data distribution only achieved a sensitivity of 90.2%, but it increased drastically to 99.7% after being integrated with the hybrid sampling technique SMOTE-ENN

[11]. This proves that without addressing class imbalance, the model carries a high risk of failing to identify patients who truly require medical treatment.

To overcome these limitations, approaches using SMOTE-Variants have become a more effective solution compared to standard oversampling methods. Variations such as SMOTE-Tomek and SMOTE-ENN have been proven capable of provid-ing better average Area Under Curve (AUC) and recall scores in handling data imbalance [12]. These techniques work not only by generating new synthetic samples but also by cleaning noise at the class boundaries (decision boundary), thereby improving the model's ability to distinguish between majority and minority classes. Therefore, this study proposes a com-parison of various SMOTE variations combined with feature selection optimization and hyperparameter tuning based on Particle Swarm Optimization in the Decision Tree algorithm. Through this approach, it is expected to produce a breast cancer prediction model that is not only globally accurate but also has high sensitivity in detecting the minority class.

II. METHODOLOGY

A. Research Workflows

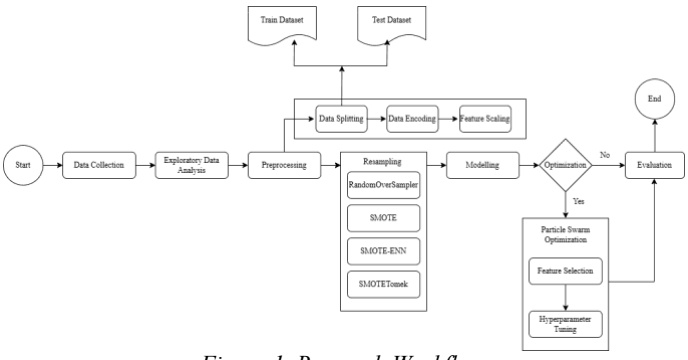


Figure 1. Research Workflows

The research methodology follows a structured workflow illustrated in Fig 1. The process begins with collecting the breast cancer dataset, followed by Exploratory Data Analysis (EDA) to understand the characteristics of the data. The data preprocessing stage is then carried out systematically, including data splitting, data encoding, and feature scaling to normalize attribute values. After the data is prepared, the resampling stage is applied using four techniques, namely RandomOverSampler, SMOTE, SMOTE-ENN, and SMOTE-Tomek, to address the class imbalance problem.

The balanced data is then processed in the modeling stage using the Decision Tree algorithm. To achieve optimal re-sults, an optimization pathway is implemented using Particle Swarm Optimization (PSO), which performs a dual role as a feature selection and hyperparameter tuning instrument. The final stage of the study is a comprehensive evaluation that not only measures predictive performance through accuracy, precision, recall, and F1-score, but also assesses the compu-tational efficiency of the model. This efficiency measurement is conducted by analyzing the average latency (Avg Latency) and throughput to compare the effectiveness of each SMOTE variation on the optimized Decision Tree model.

B. Data Collection

Data collection is a crucial stage that determines the validity of performance comparisons in this study. The researchers used the Breast Cancer Dataset from the UCI Machine Learning Repository, which consists of clinical data from the University of Wisconsin Hospitals, Madison, contributed by Dr. William H. Wolberg [13]. The selection of this dataset was intentionally made to maintain consistency with previous stud-ies [9], so that performance comparisons between the proposed model and earlier methods could be conducted accurately and objectively. A deep understanding of the characteristics of the attributes in this dataset provides a strong foundation for identifying performance gaps, particularly in the aspect of handling data imbalance, which had not been optimally addressed in previous studies.

C. Exploratory Data Analysis

The Exploratory Data Analysis (EDA) stage was conducted to understand the characteristics of the dataset through feature correlation visualization and class label distribution analysis. This step aims to identify data redundancy and the level of class imbalance as a foundation for the resampling and optimization processes in the subsequent stages.

D. Preprocessing

The preprocessing stage was carried out to transform raw data into a format ready to be used by the classification model in order to avoid technical issues that could affect prediction results. The first step involved Data Splitting using the Stratified K-Fold Cross Validation method with a value of k = 10 to ensure that each fold maintained a balanced proportion of class labels [14], [15]. Finally, Feature Scaling was performed using StandardScaler to normalize all attributes so that they had a uniform distribution scale [16]. The entire sequence of processes aimed to ensure that the dataset had consistent and stable quality, allowing the model to learn patterns from each feature optimally without bias caused by differences in value scales.

E. Resampling

The data balancing stage was carried out to address the class imbalance problem through four different approaches to ensure that the model was not biased toward the majority class. Random Oversampling was used as a baseline by randomly duplicating minority class samples to balance the number of data points without altering their original variation [17], [18]. The second approach was SMOTE, which works by synthesizing new samples along the line segments between minority class samples to improve model generalization [19]. Furthermore, the hybrid sampling method SMOTE-ENN was implemented to perform oversampling while simultaneously removing noisy or borderline samples using Edited Nearest Neighbor (ENN), thereby producing a cleaner and more robust data representation [11]. Lastly, the SMOTE-Tomek technique was applied by combining synthetic sample generation with the removal of closely paired data points through Tomek Links to sharpen the boundaries between overlapping classes [20].

Through the comparison of these four techniques, this study aims to identify the most effective resampling method for improving diagnostic sensitivity in the breast cancer dataset.

F. Modelling

The modeling stage in this study employed the Decision Tree algorithm, which is a model that transforms data into a hierarchical decision tree structure and logical decision rules. The main advantage of this model lies in its ability to handle complex decision-making by simplifying problems through a branching structure. In constructing a decision tree, there are two fundamental calculations used, namely Gain and Entropy [21].

Information Gain is used to determine which attribute is the most effective in classifying the data. The calculation of Gain is formulated in Equation (1):

$$Gain(S,A) = Entropy(S) - \frac{\sum_{i=1}^{n} |S_i|}{|S|} \times Entropy(S_i)$$

Where S represents the set of cases, A is the attribute, n is the number of partitions of attribute A, $|S_i|$ is the number of cases in the i-th partition, and |S| is the total number of cases in S. Meanwhile, Entropy is used to measure the level of disorder or homogeneity within a dataset, which is formulated in Equation (2):

$$Entropy(S) = \sum_{i=1}^{n} -p_i \times \log_2(p_i)$$

Where pi represents the proportion of S_i to S. Through these two calculations, the algorithm can determine the best splitting attribute at each level of the tree in order to minimize uncertainty in breast cancer tumor classification.

G. Optimization

The optimization stage was carried out to improve model performance through the search for the most optimal feature set and parameter configuration using the Particle Swarm Optimization (PSO) algorithm. PSO was applied to perform two integrated functions, namely feature selection and hyperparameter tuning for the Decision Tree algorithm [22].

In the feature selection process, PSO is used to remove redundant features through evaluation of different attribute combinations. In addition, PSO optimizes four Decision Tree hyperparameters, namely max depth, min samples split, min samples leaf, and criterion (Gini or Entropy).

Each particle in the PSO population moves iteratively to update its position based on the personal best (pbest) and global best (gbest). The particle velocity and position update processes are formulated in Equations (3) and (4):

$$v_i^{t+1} = w v_i^t + c_1 r_1 (pbest_i - x_i^t) + c_2 r_2 (gbest - x_i^t)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1}$$

Where w is the inertia coefficient, c_1 and c_2 are the cognitive and social factors, while r_1 and r_2 are random values [23]. This optimization aims to produce a breast cancer classification model with high generalization capability and optimal feature efficiency.

H. Evaluation

The evaluation stage was conducted to measure the predictive effectiveness and computational efficiency of the proposed model. Classification performance was calculated based on the values of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [24]. The metrics of accuracy, precision, recall, and F1-score are respectively formulated in Equations (5)–(8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

In addition to classification performance, this study evaluated efficiency aspects to ensure that the model can operate optimally on devices with limited resources. The efficiency parameters, including average latency (Avg Latency) and throughput, are formulated in Equations (9)–(10).

$$Latency(ms) = \frac{T_{end} - T_{start}}{N} \times 1000$$

$$Troughput = \frac{1000}{Avg\ Latency}$$

Where $T_{end} - T_{start}$ is the total execution time in one fold, N is the number of test samples, and 1000 is the conversion factor to milliseconds [25]. This comprehensive evaluation aims to identify the model configuration that provides the best balance between diagnostic accuracy and data processing speed.

III. Experimental Setup

This study was implemented using the Python 3.12 programming language within a development environment supported by popular libraries such as Scikit-learn for machine learning algorithms, PySwarm for optimization, and Pandas and NumPy for data manipulation. All experiments were conducted on hardware specifications consisting of an Intel Core i5-8350U CPU, 8GB RAM, and the Windows 11 operating system.

IV. Result And Discussion

A. Dataset

This study used the breast cancer dataset from the UCI Machine Learning Repository, which consists of 30 numerical input features extracted from digital images of fine needle aspirate (FNA) samples and one diagnosis label as the target variable [26]. Technically, this dataset includes ten fundamental features: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, each measured in three different statistical dimen-sions. The number 1 in the feature name represents the mean value, the number 2 indicates the standard error value, and the number 3 represents the largest or worst value detected in the sample [27]. Complete details regarding the feature names and data types are presented in Table I.

Table 1. Features In The Breast Cancer Dataset

No	Features	Type
1	*radius1*	Float

2	*texture1*	Float
3	*perimeter1*	Float
4	*area1*	Float
5	*smoothness1*	Float
6	*compactness1*	Float
7	*concavity1*	Float
8	*concave points1*	Float
9	*symmetry1*	Float
10	*fractal dimension1*	Float
11	*radius2*	Float
12	*texture2*	Float
13	*perimeter2*	Float
14	*area2*	Float
15	*smoothness2*	Float
16	*compactness2*	Float
17	*concavity2*	Float
18	*concave points2*	Float
19	*symmetry2*	Float
20	*fractal dimension2*	Float
21	*radius3*	Float
22	*texture3*	Float
23	*perimeter3*	Float
24	*area3*	Float
25	*smoothness3*	Float
26	*compactness3*	Float
27	*concavity3*	Float
28	*concave points3*	Float
29	*symmetry3*	Float
30	*fractal dimension3*	Float
31	*Diagnosis*	Binary

B. *Exploratory Data Analysis*

Exploratory data analysis was conducted to understand the feature distribution and relationships among variables before entering the modeling stage.

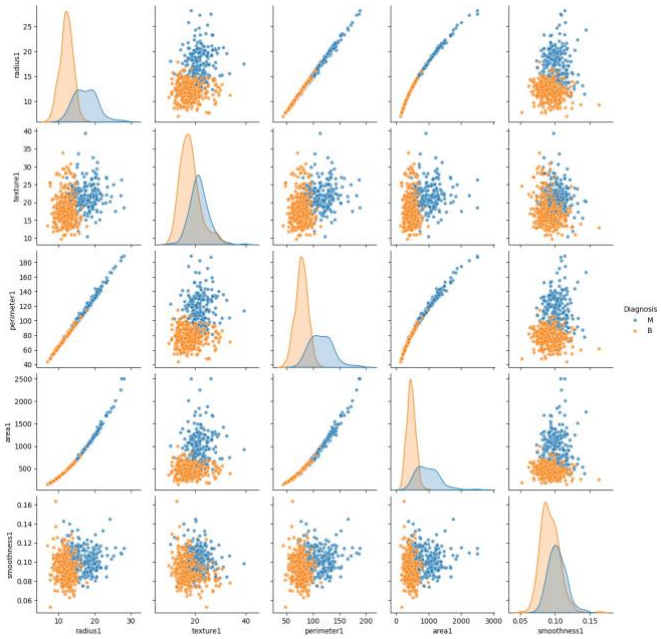


Figure 2. Pairplot of Mean Features Colored by Diagnosis

Based on the scatter matrix visualization in Fig 2, a strong linear correlation was observed among physical dimension features such as radius1, perimeter1, and area1. This indicates data redundancy, where these features provide similar geomet-ric information related to the target variable. In addition, the density plots along the diagonal in Fig 2 show that feature values for Malignant samples tend to be higher and exhibit a wider distribution compared to Benign samples.

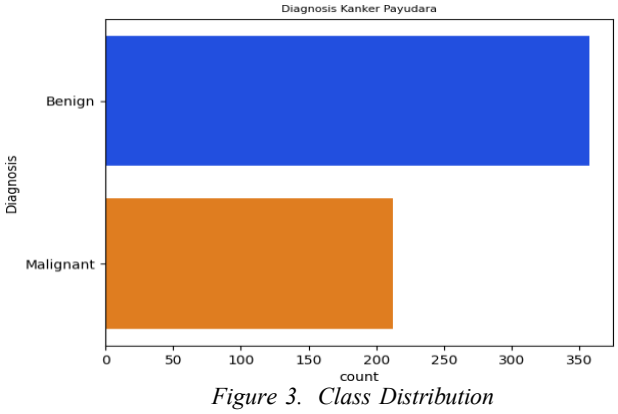


Figure 3. Class Distribution

Next, an examination of the class label distribution was conducted to identify potential data imbalance. Observations in Fig 3 show that the number of samples in the Benign category dominates compared to the Malignant category. This class distribution imbalance serves as the main rationale for applying resampling techniques in the preprocessing stage to prevent model bias toward the majority class and to ensure objective classification performance for both diagnostic categories.

C. *Preprocessing and Resampling Analysis*

This stage evaluates the transformation of raw data into an optimal format for the classification model. The primary focus is the effectiveness of Feature Scaling using StandardScaler, where the comparison of feature distributions before and after scaling is presented in Fig 4.

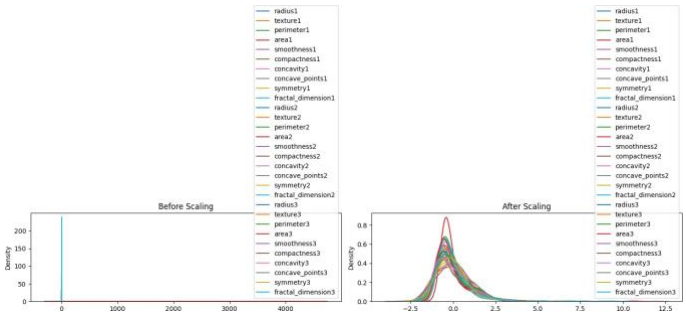


Figure 4. Feature Distributions Before and After Feature Scaling

Based on Fig 4, the features before scaling exhibit highly contrasting value ranges, such as the *area* feature with magnitudes in the thousands compared to the *smoothness* feature, which is below 1. This transformation successfully normalizes all feature scales to a mean of zero and a standard deviation of one, thereby preventing algorithmic bias toward high-magnitude features and accelerating optimization convergence. Next, an analysis was conducted on the handling of class imbalance using four resampling methods. The comparison of sample sizes for each method is presented in Table II.

Table 2. Comparison of the Number of Samples After Resampling

Method	Benign	Malignant	Total

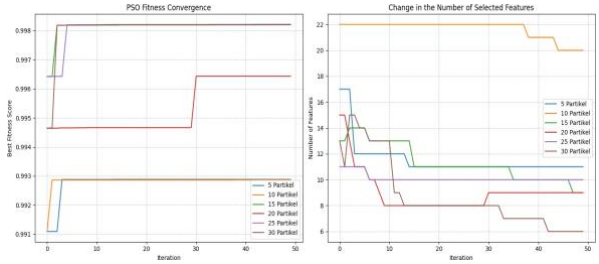
Original	357	212	569
Random Oversampling	357	357	714
SMOTE	357	357	714
SMOTE-ENN	314	307	621
SMOTE-Tomek	348	348	696

Based on Table II, both Random Oversampling and SMOTE produce identical class distributions through data duplication and synthesis. However, hybrid methods such as SMOTE-ENN and SMOTE-Tomek result in a lower total number of samples due to the integration of undersampling steps. Specifically, SMOTE-ENN shows the most significant reduction in samples because it aggressively removes observations considered as noise or located in overlapping regions. In contrast, SMOTE-Tomek performs a more selective cleaning by only removing pairs of data points from different classes that are close to each other (Tomek Links), resulting in a larger dataset while still producing clearer class boundaries.

D. Optimization Analysis

The optimization stage was conducted to identify the most discriminative combination of features and model parameters to improve classification performance. This process is divided into two main stages: feature selection and hyperparameter tuning using the Particle Swarm Optimization (PSO) algorithm.

1. Feature Selection Result. The PSO implementation for feature selection uses a fitness function based on accuracy with a penalty on the number of features to encourage model efficiency. Based on the experimental results, the configuration with 30 particles is the best-performing scenario, as it achieves the most optimal and efficient solution. PSO successfully reduces the data dimensionality from 30 original features to only 6 selected features, *namely texture1, symmetry1, radius2, area3, smoothness3, and symmetry3*.



Visualization in Fig 5 shows a sensitivity analysis of different particle numbers. The fitness convergence plot (left) demonstrates that using 30 particles (brown line) achieves highly stable fitness scores (0.9982). Meanwhile, the feature count plot (right) shows that 30 particles provide stronger exploration capability, reducing feature redundancy and leaving only 6 main attributes at the final iteration. This 80% reduction in the total number of features indicates that the model can operate efficiently using only the most representative texture, symmetry, and geometric information related to tumor malignancy.

2. Hyperparameter Tuning Result. After selecting the best features, PSO was applied again to optimize four Decision Tree hyperparameters. The search space is defined for max depth, min samples split, min samples leaf, and the criterion selection (Gini or Entropy). The final results of the parameter tuning using 20 particles over 30 iterations are presented in Table III.

Table 3. PSO-Based Hyperparameter Tuning Results

Hyperparameter	**Optimal Value**
max_depth	10
min_samples_split	8
min_samples_leaf	3
criterion	Gini
Best Fitness (Accuracy)	**0.998230**

The use of the Gini criterion with a maximum depth of 10 levels shows that the model is able to learn the data structure in depth while remaining controlled through the min_samples_leaf constraint to prevent overfitting.

E. Performance Comparison and Discussion

This stage presents a comprehensive performance comparison between the baseline Decision Tree (DT) model and various improved scenarios using resampling techniques and PSO optimization. The evaluation results of all models are presented in Table IV, while detailed classification visualizations for each scenario are shown through the Confusion Matrix in Fig 6.

Table 4. Comprehensive Performance Comparison of All Model Scenarios

Model Scenarios	**Acc**	**Prec**	**Rec**	**F1**	**Latency**	**Throughput**
DT (Baseline)	0,9261	0,9299	0,9261	0,9259	0,0331 ms	30.153
DT + RandomOverSampler	0,9262	0,9295	0,9262	0,9261	0,0445 ms	22.437
DT + SMOTE	0,9425	0,9444	0,9425	0,9425	0,0201 ms	49.523
DT + SMOTE-ENN	0,9887	0,9888	0,9887	0,9887	0,0555 ms	17.996
DT + SMOTE-Tomek	0,9919	0,9923	0,9919	0,9919	0,0835 ms	11.974
DT + S-Tomek + PSO (FS)	**0,9968**	**0,9968**	**0,9968**	**0,9967**	**0,0036 ms**	**274.897**
DT + S-Tomek + PSO (FS+T)	0,9968	0,9968	0,9968	0,9967	0,0124 ms	80.550

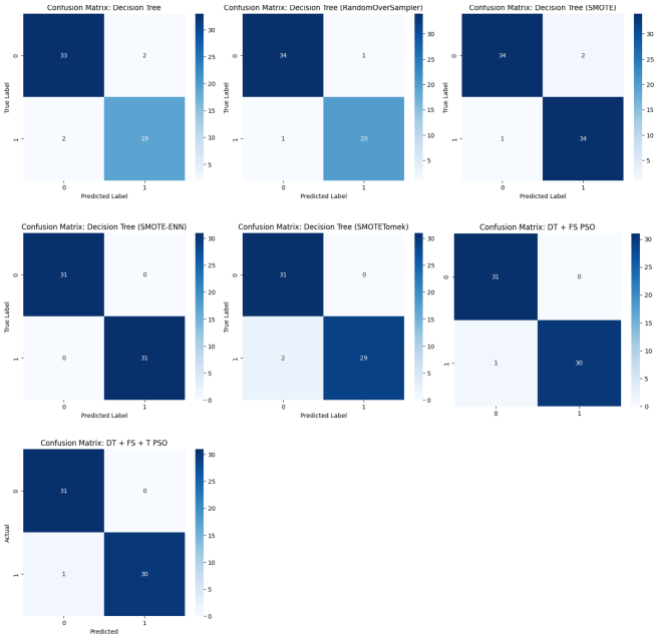


Figure 6. Confusion Matrix for Various Experimental Scenarios

Based on Table IV, there is a clear trend of performance improvement as preprocessing techniques and optimization are progressively applied. The baseline model yields the lowest accuracy (0.9261), where Fig 6 shows 4 misclassifications (2 False Positives and 2 False Negatives). The use of hybrid resampling techniques, particularly SMOTE-Tomek, proves

more effective in cleaning class boundaries, leaving only 2 False Negative errors.

The integration of PSO contributes the most significant improvement. In the DT + FS PSO scenario, misclassifications are reduced dramatically to only 1 False Negative sample. This demonstrates that PSO-based feature selection not only improves efficiency but also enhances the discrimination of essential features. The model with PSO feature selection achieves the lowest latency of 0.0036 ms per sample with a throughput of 274,897 samples per second. This is due to the reduction of features from 30 to 6 essential attributes, which directly reduces the computational burden of the decision tree structure.

Although the addition of hyperparameter tuning (FS+T) maintains similar accuracy and confusion matrix patterns as the FS scenario, latency increases to 0.0124 ms due to higher tree complexity. Overall, the combination of SMOTE-Tomek and PSO-based feature selection is the best-performing scenario, achieving near-perfect diagnostic accuracy (99.68%) with only 1 failure in the test data, as well as very high processing speed suitable for real medical implementation.

F. Feature Importance Analysis

To provide transparency in model decision-making, an inter-pretability analysis was conducted using SHAP to understand the contribution of each feature to the diagnostic predictions. The SHAP Summary Plot visualization in Fig 7 presents the distribution of feature importance across the six essential features selected by the PSO algorithm.

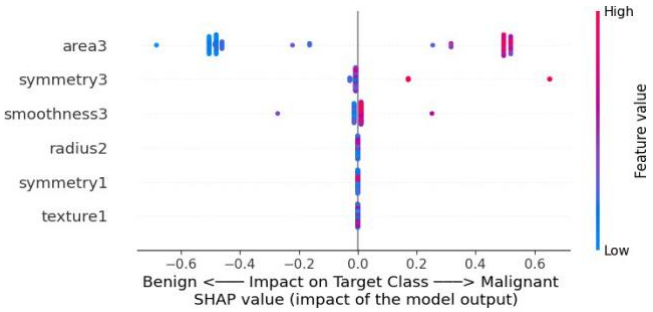


Figure 7. SHAP Summary Plot for Model Interpretation

Based on Fig 7, the feature *area3* (the largest value of the cell nucleus area) emerges as the most dominant predictor in determining tumor malignancy. High feature values (shown in red) are positively correlated with the *Malignant* classification, while low values (shown in blue) strongly indicate the *Benign* class. This insight is consistent with medical literature, where enlargement or abnormal size of the cell nucleus is a key indicator of cancer cell proliferation [28].

Other supporting features such as *symmetry3* and *smoothness3* also show significant influence, although with a more concentrated distribution. Collectively, this XAI analysis demonstrates that although the Decision Tree model operates mathematically, its decision-making process is still grounded in clinically relevant cellular biological characteristics. This provides additional confidence for medical practitioners to use this AI-based decision support system in early breast cancer diagnosis procedures.

REFERENCES

[1] A. Latifah, A. Alrizal, and Y. Fendriani, "Klasifikasi Penyakit Kanker Payudara pada Citra Mammogram Menggunakan Algoritma Convolu-tional Neural Network (CNN) dan Random Forest," *Journal of Online Physics*, vol. 10, no. 3, pp. 95–103, Jul. 2025.

[2] H. Oktavianto and R. P. Handri, "Analisis Klasifikasi Kanker Payudara Menggunakan Algoritma Naive Bayes," *INFORMAL Informatics Jour-nal*, vol. 4, no. 3, p. 117, 2020.

[3] I. N. Atthalla, A. Jovandy, and H. Habibie, "Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K-Nearest Neighbor," in *Proceedings of Annual Research Seminar*, vol. 4, no. 1, 2018, pp. 978–979.

[4] D. Kashyap, D. Pal, R. Sharma, V. K. Garg, N. Goel, D. Koundal,
A. Zaguia, S. Koundal, and A. Belay, "Global Increase in Breast Cancer Incidence: Risk Factors and Preventive Measures," *BioMed Research International*, vol. 2022, pp. 1–16, 2022.

[5] World Health Organization, "Breast Cancer," https://www.who.int/news-room/fact-sheets/detail/breast-cancer, 2024, [Online; accessed 10-Apr-2026].

[6] A. Rizka, M. K. Akbar, and N. A. Putri, "Carcinoma Mammae Sinistra T4bN2M1 Metastasis Pleura," *AVERROUS: Jurnal Kedokteran dan Kesehatan Malikussaleh*, vol. 8, no. 1, p. 23, 2022.

[7] C. Chazar and B. Erawan, "Machine Learning Diagnosis Kanker Payu-dara Menggunakan Algoritma Support Vector Machine," *Informasi: Jurnal Informatika dan Sistem Informasi*, vol. 12, no. 1, pp. 67–80, 2020.

[8] J. J. Pangaribuan and V. Angkasa, "Komparasi Tingkat Akurasi Random Forest dan KNN untuk Mendiagnosis Penyakit Kanker Payudara," *Journal Information System Development (ISD)*, vol. 7, no. 1, Jan. 2022.

[9] J. O. Afolayan, M. O. Adebiyi, M. O. Arowolo, C. Chakraborty, and
A. A. Adebiyi, "Breast Cancer Detection Using Particle Swarm Opti-mization and Decision Tree Machine Learning Technique," in *Intelligent Healthcare*. Springer, Jun. 2022, pp. 61–83.

[10] C. Alna and C. R. Gunawan, "Pendekatan Machine Learning untuk Meningkatkan Akurasi Klasifikasi pada Dataset Tidak Seimbang," *JID (Jurnal Info Digit)*, vol. 2, no. 3, Sep. 2024.

[11] M. A. Latief, L. R. Nabila, W. Miftakhurrahman, S. Ma'rufatullah, and H. Tantyoko, "Handling Imbalance Data Using Hybrid Sampling SMOTE-ENN in Lung Cancer Classification," *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 3, no. 1,
pp. 11–18, Mar. 2024.

[12] M. Rahman, "Perbandingan SMOTE-Variants untuk Mengatasi Ketidak-seimbangan Data pada Prediksi Cacat Software," Banjarbaru, Nov. 2023.

[13] T. O. Omotehinwa, D. O. Oyewola, and E. G. Dada, "A Light Gradient-Boosting Machine Algorithm with Tree-Structured Parzen Estimator for Breast Cancer Diagnosis," *Healthcare Analytics*, vol. 4, p. 100218, Dec. 2023.

[14] S. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-Fold Cross Validation Optimization on Machine Learning for Prediction," *Sinkron*, vol. 6, no. 4, pp. 2407–2414, Oct. 2022.

[15] Wijiyanto, A. I. Pradana, Sopingi, and V. Atina, "Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa," *Jurnal Algoritma*, vol. 21, no. 1, 2023. [Online]. Available: https://jurnal.itg.ac.id/index.php/algoritma

[16] M. A. S. Aritonang, M. J. Simanulang, T. P. Batubara, I. Zega, and
M. H. Afrizal, "Natural Language Processing (NLP) and Support Vector Machine (SVM) Optimization in Detecting Phishing Website URLs," *Jurnal Teknik Informatika (JUTIF)*, vol. 7, no. 1, pp. 552–570, Feb. 2026.

[17] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Tech-niques," *IEEE Access*, vol. 8, pp. 67 899–67 911, Apr. 2020.

[18] H. Ding, L. Chen, D. Liang, Z. Fu, and X. Cui, "Imbalanced Data Classification: A KNN and Generative Adversarial Networks-Based Hybrid Approach for Intrusion Detection," *Future Generation Computer Systems*, vol. 131, p. 7, Feb. 2022.

[19] F. Li, W. Ma, H. Li, and J. Li, "Improving Intrusion Detection System Using Ensemble Methods and Over-Sampling Technique," in *Proceed-ings of the 4th International Academic Exchange Conference on Science and Technology Innovation (IAECST)*, Dec. 2022.

[20] D. S. Assyifa and A. Luthfiarta, "SMOTE-Tomek Re-sampling Based on Random Forest Method to Overcome Unbalanced Data for Multi-class Classification," *Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 2, p. 151, Jul. 2024.

[21] A. S. Biyantoro and B. Prasetiyo, "Application of Decision Tree for Health Status Classification, Compared to KNN and Naive Bayes," *IJIRSE: Indonesian Journal of Informatic Research and Software Engi-neering*, vol. 4, no. 1, pp. 47–55, Mar. 2024.

[22] B. P. Lohani, A. Dagur, and D. K. Shukla, "An Efficient Approach for Diabetes Classification Using Feature Selection and Hyperparameter Tuning," *Recent Advances in Electrical & Electronic Engineering*, vol. 18, no. 7, pp. 793–807, Aug. 2025.

[23] I. P. A. W. W. Putra, I. P. G. H. Suputra, and I. B. G. Sarasvananda, "Optimasi Hyperparameter CART Menggunakan Particle Swarm Op-timization (PSO) untuk Klasifikasi Penyakit Stroke," *JNATIA*, vol. 4, no. 1, pp. 161–170, Nov. 2025.

[24] A. Widiyanto, M. Prameswari, and M. A. Latief, "Gambling Comments Detection on YouTube: A Comparative Study of Tree-Based Boosting, LSTM and GRU Models," *JUTI: Jurnal Ilmiah Teknologi Informasi*, vol. 23, no. 2, Jul. 2025.

[25] F. H. Putra, A. R. Albar, A. S. N. Akbar, A. Kurniawan, L. Okta, F. Fitriah, and M. Muntahanah, "Analysis of the Effectiveness of Voice Command Recognition Using Machine Learning Algorithms to Support Digital Learning in the Bengkulu Region," *MESTAKA: Jurnal Pengabdian Kepada Masyarakat*, vol. 4, no. 4, pp. 452–459, Aug. 2025.

[26] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1993.

[27] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear Feature Extraction for Breast Tumor Diagnosis," in *Proceedings of SPIE: Electronic Imaging*, Jul. 1993.

[28] P. Michalakis, D. Vasilaki, A. J. Abdallah, C. Asikis, A. Niakou, A. Stratos, A. Tsouknidas, E. Johnstone, and K. Michalakis, "An Insight into Cancer Cells and Disease Progression Through the Lens of Mathematical Modeling," *Current Issues in Molecular Biology*, vol. 47, no. 7, p. 477, Jun. 2025.